Nucleic Acids Research, 2004, Vol. 32, Database issue ***D203–D207***
DOI: 10.1093/nar/gkh027

HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database

Lucy A. Stebbings* and Kenji Mizuguchi

Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

Received August 15, 2003; Revised August 22, 2003; Accepted September 3, 2003

ABSTRACT

HOMSTRAD (http://www-cryst.bioc.cam.ac.uk/homstrad/) is a collection of protein families, clustered on the basis of sequence and structural similarity. The database is unique in that the protein family sequence alignments have been specially annotated using the program, JOY, to highlight a wide range of structural features. Such data are useful for identifying key structurally conserved residues within the families. Superpositions of the structures within each family are also available and a sensitive structure-aided search engine, FUGUE, can be used to search the database for matches to a query protein sequence. Historically, HOMSTRAD families were generated using several key pieces of software, including COMPARER and MNYFIT, and held in a number of flat files and indexes. A new relational database version of HOMSTRAD, HOMSTRAD BETA (http://www-cryst.bioc.cam.ac.uk/homstradbeta/) is being developed using MySQL. This relational data structure provides more flexibility for future developments, reduces update times and makes data more easily accessible. Consequently it has been possible to add a number of new web features including a custom alignment facility. Altogether, this makes HOMSTRAD and its new BETA version, an excellent resource both for comparative modelling and for identifying distant sequence/structure similarities between proteins.

INTRODUCTION

When examining proteins, information relating to the evolutionary origin and to the function and mode of action of a protein are held in both their sequence and their structure. If function is to be maintained, structure is likely to be more rigorously conserved than sequence, so this information can be particularly valuable when comparing distantly related proteins. Additionally, clustering proteins into families based on both criteria and observing the conservation of residues/structural elements and their positions within a protein, can highlight aspects that are characteristic of a family. We provide a database that allows both sequence and structure relationships between homologous proteins to be examined simultaneously.

HOMSTRAD (1,2), together with the structural annotation program, JOY (3), and the homology recognition software, FUGUE (4), have proved effective tools for use in expert led and automated comparative modelling (5 and http://predictioncenter.llnl.gov/casp5/Casp5.html, http://www.cs.bgu.ac.il/~dfischer/CAFASP3), in the characterization of protein families (6,7) and in the identification of previously unrecognized similarities between proteins. The database has also been used to prepare environment-specific substitution tables (8,4) and for benchmarking other data sets (9,10).

HOMSTRAD provides combined protein sequence and structure information extracted from the PDB (Protein Data Bank) (11), a primary protein structure repository, and relies heavily on other databases, especially Pfam (sequence-based families) (12) and SCOP (structure-based families) (13). Pfam contains carefully compiled homologous protein sequence families and is based primarily around data from SWISS-PROT (14), an annotated sequence repository. SCOP is a hierarchical classification database of protein structures from the PDB and uses expert led compilation procedures. Unlike other sequence/structure-based databases such as CATH (15), HOMSTRAD has some unique features that facilitate the comparison of homologous proteins at the sequence and structure levels. These include JOY annotated alignments (3), so that structural information can be seen in the context of a sequence alignment, and superpositions of the structures within each family.

The expanding collection of protein structure information released through the PDB has meant that a restructuring of HOMSTRAD is required in order to keep pace with family updates. Consequently, the core family information has been put into a relational database (MySQL). This new version of HOMSTRAD, referred to as HOMSTRAD BETA in this article, is currently being tested.

The revised data structures provided by HOMSTRAD BETA allow for additional information to be made available to database users, reduce data redundancy and update times, make it easier to perform complex data analyses and provide the flexibility necessary for future developments. Currently the database holds ~2700 families, just under half of which are multi-member (i.e. include more than one structure from the PDB).

*To whom correspondence should be addressed. Tel: +44 1223 760469; Fax: +44 1223 766002; Email: homstrad@cryst.bioc.cam.ac.uk

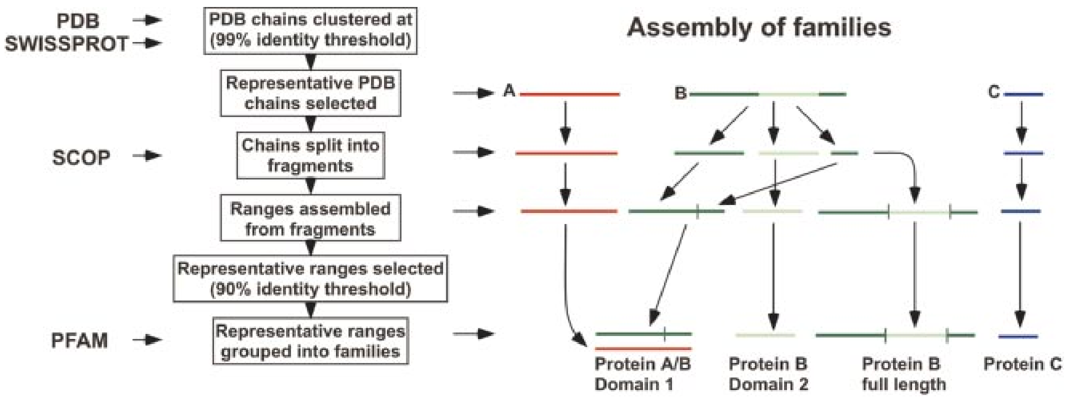

PDB
SWISSPROT
SCOP
PFAM
PDB chains clustered at (99% identity threshold)
Representative PDB chains selected
Chains split into fragments
Ranges assembled from fragments
Representative ranges selected (90% identity threshold)
Representative ranges grouped into families
Assembly of families
A
B
C
Protein A/B Domain 1
Protein B Domain 2
Protein B full length
Protein C

Figure 1. A number of steps are required to incorporate data from the PDB into HOMSTRAD BETA. First, each PDB entry is separated into chains and those containing nucleic acid data only, Cα data only, theoretical structures or data that have been discredited in some way are added to a separate data table and processed no further. Those that survive this filtering process are passed through the procedures detailed in the figure (boxed, to the left) with input from SWISS-PROT, SCOP and Pfam data at various stages as indicated (very left hand side). To the right is a schematic representation tracing the processes undergone by three sample PDB chains, one of which (chain B) encodes a hypothetical protein containing two domains, the first domain of which consists of two fragments. Families are generated corresponding to each of the two domains and also to the full-length protein. The PDB chain A sequence is homologous to domain 1 of chain B and so is incorporated into the same family. Many families in HOMSTRAD are simpler than this and at present just over 1000 include only one chain with one fragment, as is seen for the hypothetical family that contains chain C. These families will increase in size as more structures are released into the PDB and are matched to the families.

DATABASE TABLE DESIGN

In order to design the new database, an analysis of the components of HOMSTRAD was carried out, breaking it down into its smallest parts (Fig. 1). Some of these levels of division are obvious, such as the division of a family into PDB chains, whilst others are less clear. Occasionally the complexity of protein domain structures means that one domain may consist of more than one non-contiguous stretch of a PDB chain or chains. We refer to the entire region of a PDB entry that occurs in a HOMSTRAD family as a range and call each contiguous section of that range a fragment. We use the term ‘range’ rather than ‘domain’ because, although many HOMSTRAD families consist of single domains (where the definitions of domains are guided by Pfam and SCOP), we also hold multi-domain families to allow the relationships between domains to be examined. Each range in a family, therefore, corresponds to one or more domains. A particular range can belong to one and only one family but more than one overlapping or non-overlapping range from the same PDB chain can exist in the database. As will be discussed later, this range can also be a representative for multiple other highly similar ranges whose data are also stored.

We have applied standard principals of relational database design to this information. The core data for the new HOMSTRAD BETA database (held in MySQL) now has tables for PDB entries, PDB chains, representative ranges, ranges, fragments, linking information to external databases and various other tables required to maintain workable relationships.

FAMILY GENERATION PROCEDURE

In the core HOMSTRAD BETA database there is a general flow of data in, from the PDB (this is clustered to give a non-redundant data set—see next section), SWISS-PROT (14), Pfam (12) and SCOP (13), which converges at the ‘range’ level (Fig. 1). This assorted text and sequence information from various sources is then integrated and the ranges are clustered into HOMSTRAD families using BLAST sequence searches and clustering information from Pfam. At this point manual input is required to check the families and make any alterations that are needed. We do not maintain a specific cut-off threshold for inclusion into a family although we have always defined a family as a group of proteins that show clear evidence of common ancestry. This is in contrast to CAMPASS (16), which includes super-family alignments of more distantly related structures. To define families in practice, in line with other partially automated databases, the quality of the alignment and superposition are judged by eye and dissimilar structures or sequences are moved into a separate family (see the family example below). When the family details are ready, the programs described in the software section below are run to generate the various files needed to display the family on the web.

CLUSTERING OF PDB CHAINS AND IDENTIFICATION OF REPRESENTATIVE STRUCTURES

Revisions have been made to the HOMSTRAD procedures used to cluster highly similar PDB protein sequences and there are now two phases to this process.

During the first stage, PDB chains are clustered into groups, each representing a distinct native protein. Initially, SWISS-PROT (14) entries are matched to corresponding PDB chains through BLAST sequence searching. Comparison of text information is used to refine these hits. Next the sequences are clustered at a 99% identity threshold and textual information from the PDB entries and the corresponding SWISS-PROT links are used to adjust the threshold and either merge clusters or split them. This results in a set of ~7500 clusters. Whilst not a totally foolproof technique this significantly improves the reliability of clustering of PDB sequences that contain disordered or mutated regions so that less manual adjustments have to be made later in the process.

The second clustering phase is carried out as the families are compiled from ranges. This step is sequence based and a threshold of 90% identity is used, i.e. no two representative ranges within a family should have an identity higher than 90%. Without this stage, many highly similar proteins could be present in a family, obscuring relationships between less similar proteins and skewing the data content, potentially affecting the profiles used by FUGUE (4).

The comprehensive SWISS-PROT links are a key improvement in HOMSTRAD BETA and have uses outside clustering methods. Whilst other databases maintain reasonable SWISS-PROT/PDB links, they are difficult to keep up to date and are often defined only at the PDB entry level rather than at the PDB chain level. For protein family analysis purposes, it is proving useful to hold comprehensive information on SWISS-PROT/PDB mapping to provide a bridge between structural and sequence information in other databases.

SOFTWARE PROCEDURES USED TO GENERATE THE BETA FAMILIES

The underlying principals for generating family-specific data have remained the same. Homologous protein structures are superimposed, and an alignment produced using COMPARER (17) followed by a revised superposition using MNYFIT. The alignment is then annotated using JOY (3) and a family profile is generated for use with the FUGUE (4) search engine. Finally the family extensions are added. Evolutionary trace data are provided by TRACE SUIT II (18), extended CLUSTALW (19) sequence alignments are appended and PROSITE (20) motifs are incorporated into the alignments (2). The step used to generate the structure-based alignment is now modular so that other software can be interchanged to some extent. We are currently testing a new version of COMPARER in this role.

IMPROVEMENTS TO THE WEB-ACCESSIBLE DATABASE

The transfer of core data to a more flexible, easily queried relational format and the addition of new data, have allowed several new features to be added to the HOMSTRAD BETA web pages (Fig. 2).

Keyword search facilities have been improved allowing several different types of accession number to be input as queries.

Connections between related but distinct families are now available via the 'links to other Homstrad families' link on each family page. This provides a list of PDB sequences within the current family and their percentage identities and BLAST E-value scores to sequences in other families.

One limitation of the original HOMSTRAD database was its poor links between PDB chains and SWISS-PROT entries. As discussed earlier, this problem is being remedied in HOMSTRAD BETA and more comprehensive SWISS-PROT links are provided. Links to the corresponding Pfam families are also maintained more rigorously.

A major addition to the database is the custom alignment facility, accessed via the 'show related PDB structures and sequences' link on every family page. In the current version of HOMSTRAD, each entry within a family is a representative of a number of other PDB chains and is selected on the basis of several criteria—primarily the resolution of the structure. This does not allow for non-representative structures, which are perhaps in different complexes or have critical mutations, to be viewed. Using the custom alignment facility, different combinations of PDB entry sequence can be aligned on the fly and a JOY annotated alignment and structural superposition provided. It is also possible to add in SWISS-PROT sequences to the alignment so that any gaps in the sequence/structure information, disordered regions or mutations can be identified more easily. This kind of facility is only possible now that data on all the non-representative PDB chains are being maintained within the relational database.

AN EXAMPLE: THE GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE FAMILIES

GAPDH enzymes reversibly catalyse the oxidation and phosphorylation of D-glyceraldehyde-3-phosphate to 1,3-diphospho-glycerate, thus fulfilling an important role in glycolysis and gluconeogenesis. Several different types have been described in a wide range of organisms.

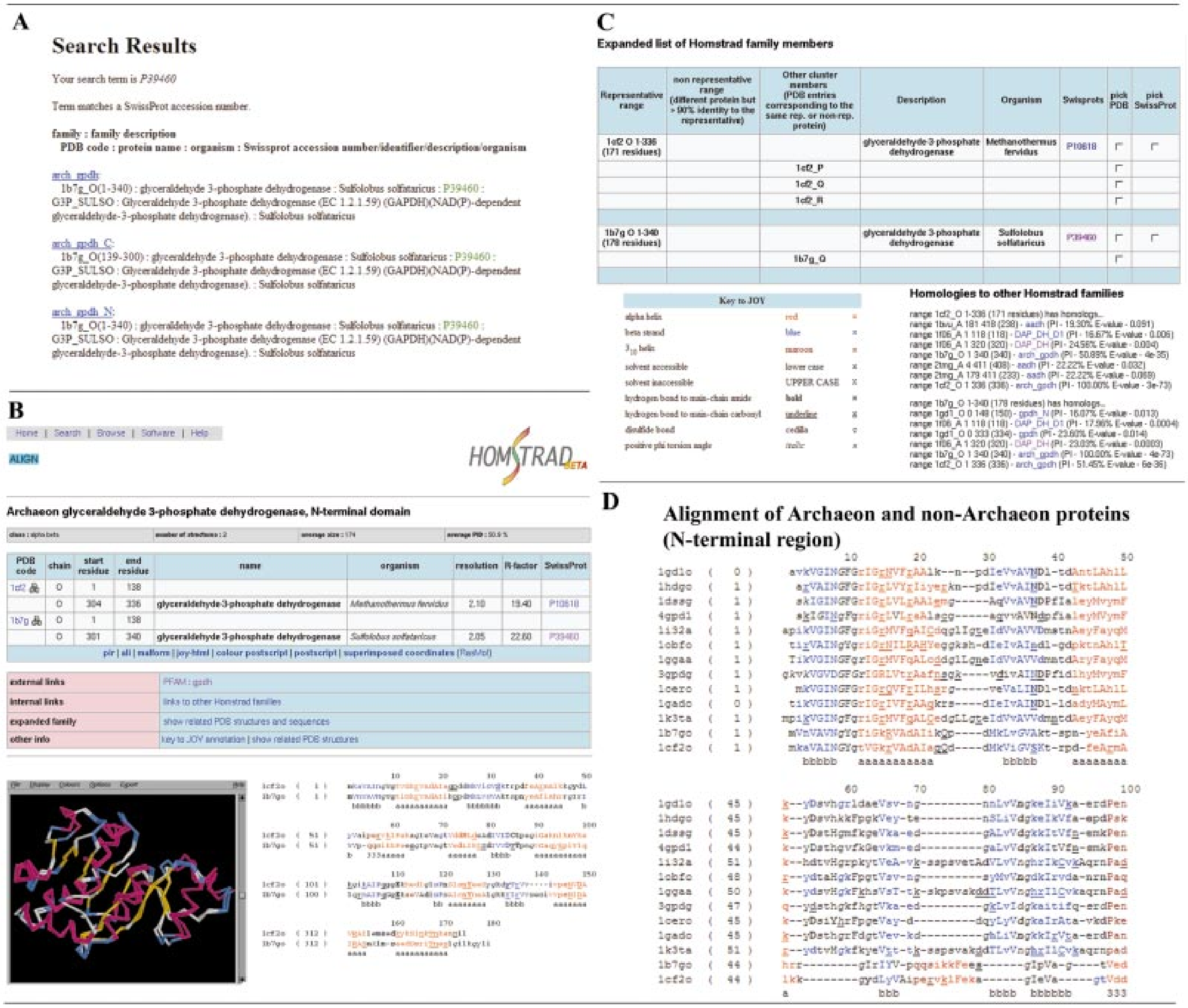
SCOP defines two domains, a NAD(P)-binding Rossmann fold domain which includes the N-terminal region plus a short stretch at the C-terminus, and a catalytic domain. Pfam holds a family for each of these domains, gpdh and gpdh_C. Both databases define quite large families, including many members from various organisms.

In HOMSTRAD, the original gpdh family was not divided into domains. HOMSTRAD BETA, however, holds both domains plus the full-length family so that the relationship between the domains can be viewed. Each of the families have also been split into two to reflect a distinct group of proteins from hyperthermophilic Archaeon (PDB entries 1cf2_O and 1b7g_O). The Archaeon protein structures diverge from GAPDH enzymes in other organisms in a number of the loop regions. Also, BLAST hits between Archaeon and non-Archaeon family members have E-value scores no lower than 0.0005 and all but one have scores above 0.01, i.e. much less significant than the scores between members within the families. Their sequence and structural differences are evident in a JOY annotated alignment that includes sequences from Archaeon and non-Archaeon families, for example in the region between around residues 50–100 (Fig. 2D). Screen shots from one of the domain families are shown in Figure 2A–C.

SUMMARY

The facilities provided by HOMSTRAD have been extended and the core data transferred to a relational data structure, HOMSTRAD BETA, which is easier to maintain and query than the previous database. Currently, HOMSTRAD and HOMSTRAD BETA run in parallel. However, once local users have adapted to using the new system we intend to transfer all operations to the relational structure and merge the databases into one, retaining the name HOMSTRAD.

This relational data structure has allowed new facilities to be added to the web pages and has improved flexibility so that



Search Results

Your search term is *P39460*

Term matches a SwissProt accession number.

family : family description
PDB code : protein name : organism : Swissprot accession number/identifier/description/organism

arch_gpdh:
1b7g_O(1-340) : glyceraldehyde 3-phosphate dehydrogenase : Sulfolobus solfataricus : P39460 : G3P_SULSO : Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.59) (GAPDH)(NAD(P)-dependent glyceraldehyde-3-phosphate dehydrogenase). : Sulfolobus solfataricus

arch_gpdh_C:
1b7g_O(139-300) : glyceraldehyde 3-phosphate dehydrogenase : Sulfolobus solfataricus : P39460 : G3P_SULSO : Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.59) (GAPDH)(NAD(P)-dependent glyceraldehyde-3-phosphate dehydrogenase). : Sulfolobus solfataricus

arch_gpdh_N:
1b7g_O(1-340) : glyceraldehyde 3-phosphate dehydrogenase : Sulfolobus solfataricus : P39460 : G3P_SULSO : Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.59) (GAPDH)(NAD(P)-dependent glyceraldehyde-3-phosphate dehydrogenase). : Sulfolobus solfataricus

Archaeon glyceraldehyde 3-phosphate dehydrogenase, N-terminal domain

class : alpha beta	number of structures : 2	average size : 174	average PID : 50.9 %

PDB code	chain	start residue	end residue	name	organism	resolution	R-factor	SwissProt
1cf2	O	1	138					
	O	304	336	glyceraldehyde-3-phosphate dehydrogenase	*Methanothermus fervidus*	2.10	19.40	P10618
1b7g	O	1	138					
	O	301	340	glyceraldehyde 3-phosphate dehydrogenase	*Sulfolobus solfataricus*	2.05	22.60	P39460

pir | ali | mallorn | joy-html | colour postscript | postscript | superimposed coordinates (RasMol)

external links	PFAM : gpdh
internal links	links to other Homstrad families
expanded family	show related PDB structures and sequences
other info	key to JOY annotation | show related PDB structures

Expanded list of Homstrad family members

Representative range	non representative range (different protein but > 90% identity to the representative)	Other cluster members (PDB entries corresponding to the same rep. or non-rep. protein)	Description	Organism	Swissprots	pick PDB	pick SwissProt
1cf2 O 1-336 (171 residues)			glyceraldehyde-3-phosphate dehydrogenase	Methanothermus fervidus	P10618	□	□
		1cf2_P				□	
		1cf2_Q				□	
		1cf2_R				□	
1b7g O 1-340 (178 residues)			glyceraldehyde 3-phosphate dehydrogenase	Sulfolobus solfataricus	P39460	□	□
		1b7g_Q				□	

Key to JOY		
alpha helix	red	x
beta strand	blue	x
3_{10} helix	maroon	x
solvent accessible	lower case	x
solvent inaccessible	UPPER CASE	X
hydrogen bond to main-chain amide	bold	x
hydrogen bond to main-chain carbonyl	underline	x
disulfide bond	cedilla	ç
positive phi torsion angle	italic	x

Homologies to other Homstrad families

range 1cf2_O 1-336 (171 residues) has homologs...
range 1bvu_A 181 418 (238) - aadh (PI - 19.30% E-value - 0.091)
range 1f06_A 1 118 (118) - DAP_DH_D1 (PI - 16.67% E-value - 0.006)
range 1f06_A 1 320 (320) - DAP_DH (PI - 24.56% E-value - 0.004)
range 1b7g_O 1 340 (340) - arch_gpdh (PI - 50.88% E-value - 4e-35)
range 2tmg_A 4 411 (408) - aadh (PI - 22.22% E-value - 0.032)
range 2tmg_A 179 411 (233) - aadh (PI - 22.22% E-value - 0.068)
range 1cf2_O 1 336 (336) - arch_gpdh (PI - 100.00% E-value - 3e-73)

range 1b7g_O 1-340 (178 residues) has homologs...
range 1gd1_O 0 149 (150) - gpdh_N (PI - 16.07% E-value - 0.013)
range 1f06_A 1 118 (118) - DAP_DH_D1 (PI - 17.96% E-value - 0.0004)
range 1gd1_O 0 333 (334) - gpdh (PI - 23.60% E-value - 0.014)
range 1f06_A 1 320 (320) - DAP_DH (PI - 23.03% E-value - 0.0003)
range 1b7g_O 1 340 (340) - arch_gpdh (PI - 100.00% E-value - 4e-73)
range 1cf2_O 1 336 (336) - arch_gpdh (PI - 51.45% E-value - 6e-36)

Alignment of Archaeon and non-Archaeon proteins (N-terminal region)

```
                            10        20        30        40        50
1gdlo   (   0 )    avkVGINGFGrIGrNVFrAAlk--n--pdIeVvAVNDl-tdAntLAhlL
1hdgo   (   1 )     arVAINGFGrIGrLVYrIiyerkn--pdIeVvAINDl-tdTktLAhlL
1dssg   (   1 )     skIGINGFGrIGrLVLrAAlerg-----AqVvAVNDPfIaleyMvymF
4gpdl   (   1 )     skIGINgFgrIGrLVLraAlecg-----agvvAVNdpfialeyMVymF
1i32a   (   1 )   apikVGINGFgrIGrMVFgAICdqglIgteIdVvAVVDmstnAeyFayqM
1obfo   (   1 )     tirVAINgVgrIGrNILRAHVeggksh-dIeIvAIndl-gdpktnAhlT
1ggaa   (   1 )    TikVGINGFgrIGrMVFqALcddglLgneIdVvAVVdmstdAryFayqM
3gpdg   (   1 )   qkvkVGVDGFGrIGRLVtrAafnsqk----vdivAINDPfidlhyMvymF
1cero   (   1 )     mKVGINGFGrIGrQVFrILhsrg-----veVaLINDl-tdaktLAhlL
1gado   (   0 )    tiKVGINGFGrIGrIVFrAAgkrs----dIeIvAINDl-ldadyMAymL
1k3ta   (   1 )   mpiKVGINgFgrIGrMVFgALCedgLLgteIdVvAVVdmgtdAeyFAyqM
1b7go   (   1 )    mVnVAVNgYgTiGkRVAdAIikQp----dMkLvGVAkt-spn-yeAfiA
1cf2o   (   1 )    mkaVAINGYgtVGkrVAdAIagQd----dMkViGVSKt-rpd-feAgnA
                     bbbbb   aaaaaaaaaaa          bbbbb     aaaaaaaa

                            60        70        80        90       100
1gdlo   (  45 )   k--yDsvhgrldaeVsv-ng----------nnLvVngkeIiVka-erdPen
1hdgo   (  45 )   k--yDsvhkkFpgkVey-te----------nSLiVdgkeIkVfa-epdPsk
1dssg   (  45 )   k--yDstHgmfkgeVka-ed----------gALvVdgkkItVfn-enkPen
4gpdl   (  44 )   k--yDsthgvfkGevkm-ed----------gaLvVdgkkItVfn-enkPen
1i32a   (  51 )   k--hdtvHgrpkytVeA-vk-sspavetAdVLvVnghrIkCvkAqrnPad
1obfo   (  48 )   r--ydtaHgkFpgtVav-ng----------syMvVngdkIrvda-nrnPaq
1ggaa   (  50 )   k--ydsvHgkFkhsVsT-tk-skpsvakddTLvVnghrIlCvkaqrnPad
3gpdg   (  47 )   q--ydsthgkfhgtVka-ed----------gkLvIdgkaitifq-erdPen
1cero   (  45 )   k--yDsiYhrFpgeVay-d-----------dqyLyVdgkaIrAta-vkdPke
1gado   (  45 )   k--yDsthgrFdgtVev-kd----------ghLiVngkkIrVta-erdPan
1k3ta   (  51 )   r--ydtvHgkfkyeVtt-tk-sspsvakddTLvVnghrIlCvkaqrnpad
1b7go   (  44 )   hrr---------gIrIYV-pqqsikkFeeg-------gIpVa-g-----tVed
1cf2o   (  44 )   lkk---------gydLyVAipervklFeka-------gIeVa------gtVdd
                  a           bbb               bbbb   bbbbbb      333
```

Figure 2. Archaeon glyceraldehyde 3-phosphate dehydrogenases. (**A**) Shows the results of a keyword search using a SWISS-PROT accession number. When the link to arch_gpdh_N is followed, the arch_gpdh_N family (includes 1cf2_O and 1b7g_O PDB chains) home page is reached (**B**). Original features such as the JOY annotated alignment are shown (B) and, if Rasmol is installed and the RasMol link is clicked, the superimposed structures can be viewed (B). (**C**) Shows two new features: the custom alignment facility chooser page is shown, which gives an expanded list of all the PDB chains that are part of the family, including the non-representative members. These can be individually selected and a custom family generated. A new facility that shows links to other HOMSTRAD BETA families is also displayed (C). Also in (C) is the key to the JOY annotated alignments. (**D**) Shows the most N-terminal section of JOY annotated alignment that includes both Archaeon and non-Archaeon protein sequences (bottom two entries), highlighting the differences.

additional features may be integrated as the database develops. In the future we plan to hold data relating to functional residues and are also exploring the possibility of using the results from domain definition software to improve the automation of domain boundary prediction. Both sets of information can be integrated more easily into HOMSTRAD now that the database has been reformatted.

ACKNOWLEDGEMENTS

We thank David Burke for assistance in setting up the relational database. This work is funded by a Wellcome Trust Research Career Development Fellowship awarded to K.M.

REFERENCES

1. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
2. de Bakker,P.I.W., Bateman,A., Burke,D.F., Miguel,R.N., Mizuguchi,K., Shi,J., Shirai,H. and Blundell,T.L. (2001) HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics*, **17**, 748–749.
3. Mizuguchi,K., Deane,C.M., Blundell,T.L., Johnson,M.S. and Overington,J.P. (1998) JOY: protein sequence–structure representation and analysis. *Bioinformatics*, **14**, 617–623.
4. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: Sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
5. Williams,M.G., Shirai,H., Shi,J., Nagendra,H.G., Mueller,J., Mizuguchi,K., Miguel,R.N., Lovell,S.C., Innis,C.A., Deane,C.M. *et al.*

(2002) Sequence–structure homology recognition by iterative alignment refinement and comparative modeling. *Proteins*, **45** (Suppl.), 92–97.

6. Kumar,P.R., Eswaramoorthy,S., Vithayathil,P.J. and Viswamitra,M.A. (2000) The tertiary structure at 1.59 Å resolution and the proposed amino acid sequence of a family-11 xylanase from the thermophilic fungus *Paecilomyces varioti bainier*. *J. Mol. Biol.*, **295**, 581–593.
7. Perutz,M.F., Paoli,M. and Lesk,A.M. (1999) Fix L, a haemoglobin that acts as an oxygen sensor: signalling mechanism and structural basis of its homology with PAS domains. *Chem. Biol.*, **6**, R291–297.
8. Overington,J., Donnelly,D., Johnson,M.S., Sali,A. and Blundell,T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.
9. Thompson,J.D., Plewniak,F. and Poch,O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
10. Guda,C., Scheeff,E.D., Bourne,P.E. and Shindyalov,I.N. (2001) A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization. *Pac. Symp. Biocomput.*, **6**, 275–286.
11. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
12. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
13. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
14. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
15. Pearl,F.M.G., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I,, Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
16. Sowdhamini,R., Burke,D.F., Huang,J.F., Mizuguchi,K., Nagarajaram,H.A., Srinivasan,N., Steward,R.E. and Blundell,T.L. (1998) CAMPASS: a database of structurally aligned protein superfamilies. *Structure*, **6**, 1087–1094.
17. Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.
18. Innis,C.A., Shi,J. and Blundell,T.L. (2000) Evolutionary trace analysis of TGF-beta and related growth factors: implications for site-directed mutagenesis. *Protein Eng.*, **13**, 839–847.
19. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
20. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P.(2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.